Extreme Stochastic Variational Inference (ESVI): Distributed Inference for Large Scale Mixture Models

Jiong Zhang* [1] Parameswaran Raman* [2] Shihao Ji [3] Hsiang-Fu Yu [4] S.V.N Vishwanathan [4] Inderjit Dhillon [1,4] (* equal contribution)

UT Austin [1], UC Santa Cruz [2], Georgia State University [3], Amazon [4]

Scaling Variational Inference

Bottleneck to Model Parallelism: VI and SVI algorithms require all parameters to fit in a single processor requiring $O(D \times K)$ storage.

- **ESVI** provides a simultaneously data and model parallel algorithm cutting storage costs down to $O\left(\frac{D \times K}{P}\right)$ where P is the number of processors
- **ESVI**: algorithm is fully distributed, asynchronous and non-blocking.
- **ESVI** updates are stochastic w.r.t the coordinates, however the update in each coordinate is exact. Therefore each step is a guaranteed ascent.

Setting - Mixture of Exponential Families

Model:

- *Observations*: $x = \{x_1, \ldots, x_N\}$, with each $x_i \in \mathbb{R}^D$
- *Local* latent variables: $z = \{z_1, \ldots, z_N\}$, with each $z_i \in \{1, \ldots, K\}$
- *Global* latent variables: $\theta = \{\theta_1, \ldots, \theta_K\}$, $\quad \pi \in \Delta_K$ (mixing coefficients)

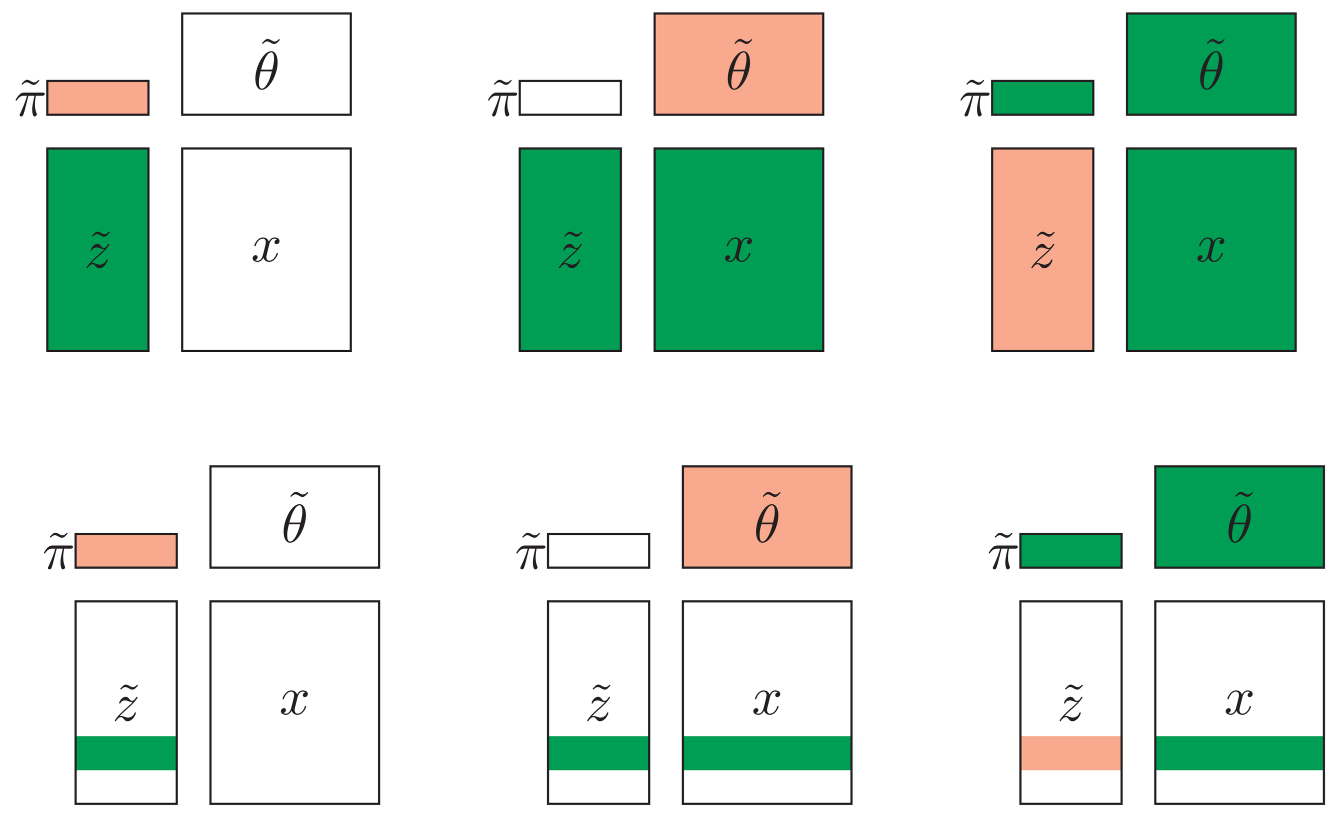
Joint Distribution:

$$p(x, \pi, z, \theta | \alpha, n, \nu) = p(\pi|\alpha) \cdot \prod_{k=1}^{K} p(\theta_k | n_k, \nu_k) \cdot \prod_{i=1}^{N} p(z_i|\pi) \cdot p(x_i|z_i, \theta)$$

Variational Distribution:

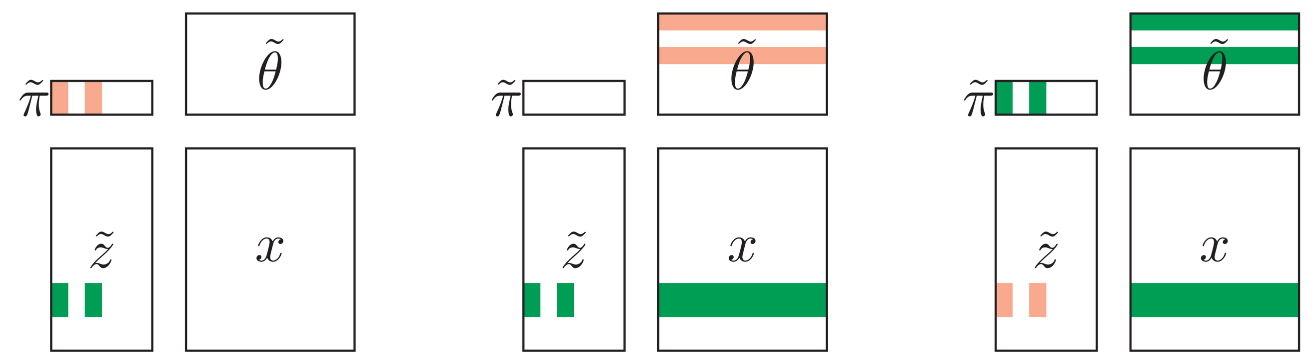
$$q\left(\pi, z, \theta | \tilde{\pi}, \tilde{z}, \tilde{\theta}\right) = q(\pi|\tilde{\pi}) \cdot \prod_{i=1}^{N} q(z_i|\tilde{z}_i) \cdot \prod_{k=1}^{K} q\left(\theta_k|\tilde{\theta}_k\right)$$

Study of Access Patterns during updates



Green indicates variable is read **Red indicates variable is updated**

ESVI



Algorithm 1 ESVI

Sample $i \in \{1, \ldots, N\}$
  Select $\mathcal{K} \subset \{1, \ldots, K\}$
    Update $\tilde{z}_{i,k}$ for all $k \in \mathcal{K}$
    Update $\tilde{\pi}_k$ for all $k \in \mathcal{K}$
    Update $\tilde{\theta}_k$ for all $k \in \mathcal{K}$

Key Idea:

1. Instead of updating all the K coordinates of the local variable $\tilde{z}_i$ and then updating all K global variables $\tilde{\pi}, \tilde{\theta}$, we only update a small subset $\mathcal{K}$ of the local variables and the corresponding global variables
2. The global variables $\tilde{\pi}$ nomadically move through the network, and this ensures mixing

Why updating a subset of K variables is a valid coordinate ascent scheme?

- Start with a feasible $\tilde{z}_i$, pick, say, a pair of coordinates $\tilde{z}_{i,k}$ and $\tilde{z}_{i,k'}$ and let $\tilde{z}_{i,k} + \tilde{z}_{i,k'} = \mathcal{C}$. If $\tilde{z}_i$ satisfied the constraints before the update, it will continue to satisfy the constraints even after the update. On the other hand, conditional ELBO increases as a result of the update.

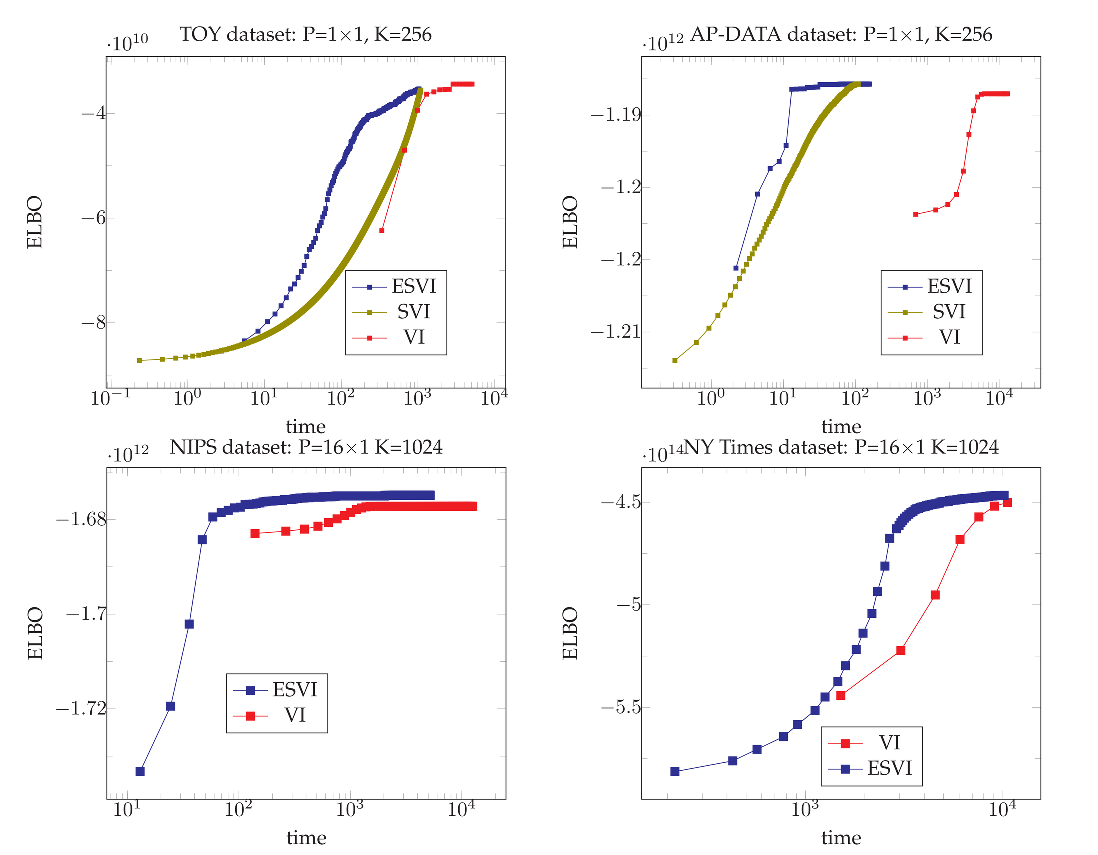
Lemma 1 *For $2 \leq K' \leq K$, let $\mathcal{K} \subset \{1, \ldots, K\}$ be s.t. $|\mathcal{K}| = K'$. For any $C > 0$, the problem*

$$\max_{z_i \in \mathbb{R}^{K'}} \mathcal{L}_{\mathcal{K}} = \sum_{k \in \mathcal{K}} \tilde{z}_{i,k} \cdot u_{i,k} - \tilde{z}_{i,k} \cdot \log \tilde{z}_{i,k} \text{ s.t. } \quad \sum_{k \in \mathcal{K}} \tilde{z}_{i,k} = C \quad \text{and} \quad 0 \leq \tilde{z}_{i,k},$$

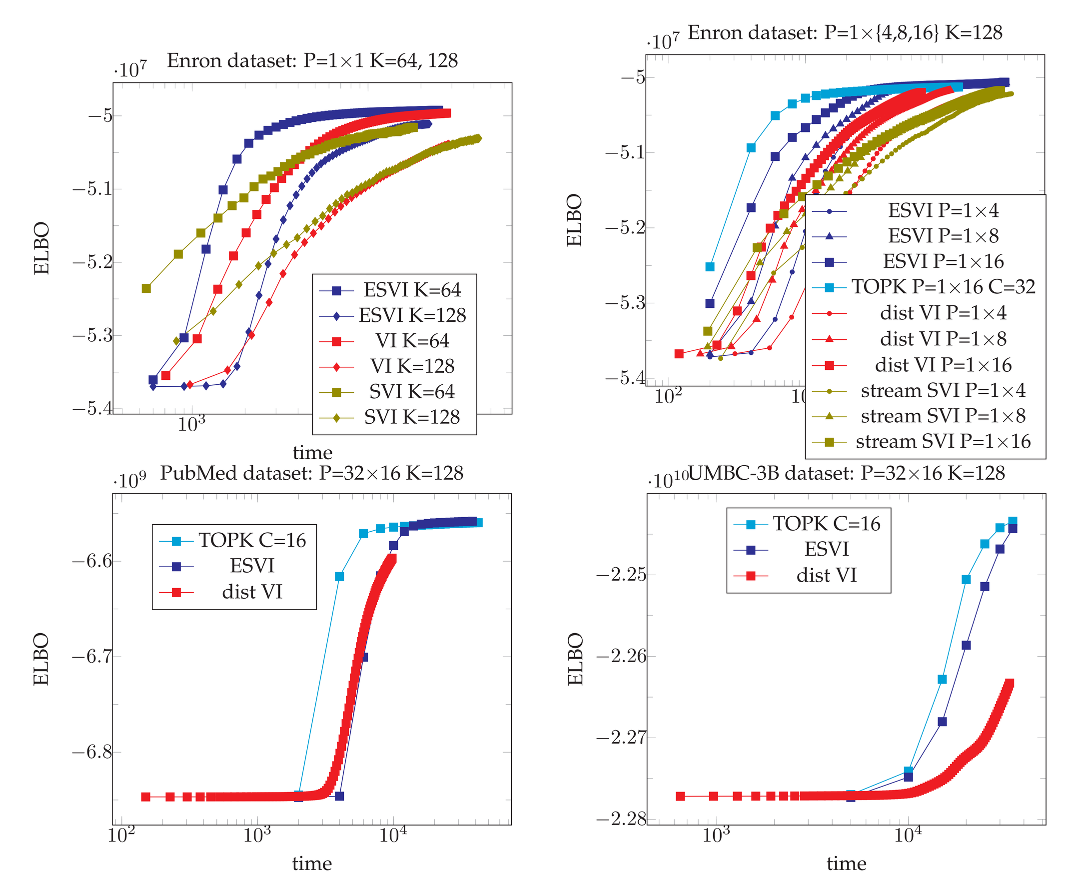
has the closed form solution:

$$\tilde{z}^*_{i,k} = C \frac{\exp(u_{i,k})}{\sum_{k' \in \mathcal{K}} \exp(u_{i,k'})}, \text{ for } k \in \mathcal{K}.$$

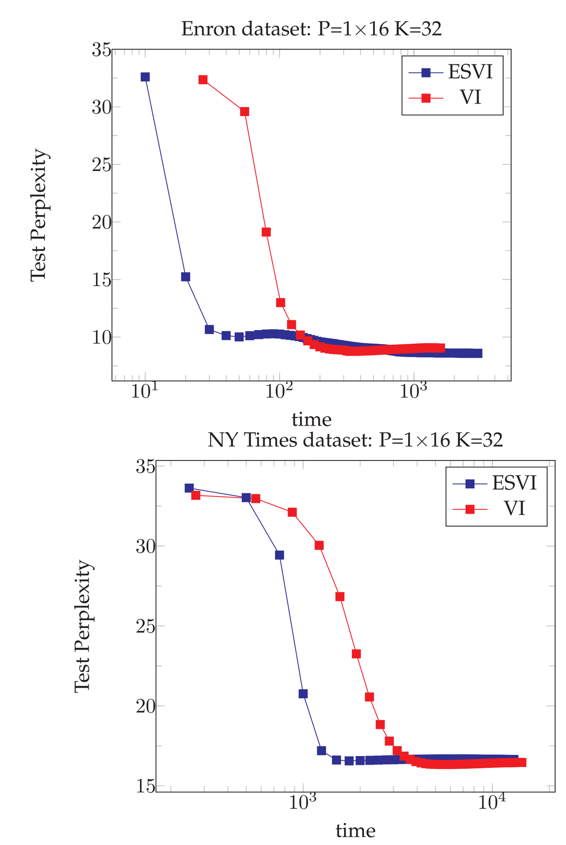
Experiments



ESVI-GMM



ESVI-LDA



Test Perplexity